The thesis of GUIDEx is that robust experimentation methods from the sciences can be adapted and applied to military experimentation and will provide the basis for advancements in military effectiveness in the transformation process.

Electronic versions of GUIDEx and its pocketbook, Slim-Ex, can be downloaded from the following site: http://www.dtic.mil/ttcp/

Guide for Understanding and Implementing Defense Experimentation

GUIDEx

Pocketbook Version of GUIDEx (Slim-Ex)

The Technical Cooperation Program

SUBCOMMITTEE ON NON-ATOMIC MILITARY RESEARCH AND DEVELOPMENT (NAMRAD)

THE TECHNICAL COOPERATION PROGRAM (TTCP)

JOINT SYSTEMS ANALYSIS (JSA) GROUP

METHODS AND APPROACHES FOR WARFIGHTING EXPERIMENTATION ACTION GROUP 12 (AG-12) (TTCP JSA AG-12)

Guide for Understanding and Implementing Defense Experimentation

GUIDEx

Pocketbook Version of GUIDEx (Slim-Ex)

Australia Canada United Kingdom United States

GUIDEx does not present the official policy of any participating nation organization. It consolidates principles and guidelines for improving the impact of science-based experimentation on defense capability development.

All organizations are invited to use the guidance it provides.

Funds for printing this document were provided by the

Canadian Forces Experimentation Centre
Ottawa, Canada

Electronic copy compiled in February 2006,
Ottawa and available at http://www.dtic.mil/ttcp/

ISBN 92-95046-12-9

Guide for Understanding and Implementing Defense Experimentation (GUIDEx)

March 2006, Version 1.1

Keywords:

experiment, trial, test, hypothesis, analysis, causal, cause-and-effect, defense, warfighting, simulation, wargame, training, exercise, ethics, valid, requirement, capability, development, measure, instrument, unit, joint, force, campaign.

Contact:
ttcp_docfeedback@dtic.mil

Art Direction by ADM(PA) DMCS CS05-0513-B

Foreword

The development of allied forces has always been a difficult and complex process. However the need for rapid force development to respond to asymmetric and unpredictable threats, the demands of coalition operations, the perceived need for information supremacy, combined with the availability of new transformational technologies and concepts, have caused this task to become even more challenging over the past few years. Experimentation offers a unique means to support the development and transformation of allied forces by advancing our knowledge of complex networked systems and capabilities likely to be fielded in the near future.

"Anything we use today arrives through a process of organized experimentation; over time, improved tools, new processes, and alternative technologies all have arisen because they have been worked out in various structured ways." (Thomke 2003: p. 1)

The importance of experimentation motivated TTCP's Joint Systems and Analysis Group (JSA) to establish Action Group 12 on *Methods and Approaches for Warfighting Experimentation* in 2002. The work of AG-12 culminated in a 350-page guide for defense experimentation - the *TTCP Guide for Understanding and Implementing Defense Experimentation (**GUIDEx**)*. GUIDEx describes **14 Principles** leading to valid (good) experimentation that are amplified through **8 Case Studies** drawn from the activities of the participating nations and coalitions. **Part I** of GUIDEx is reproduced here with appropriate additions as a standalone pocketbook on defense experimentation (known as **Slim-Ex**). Parts II and III, the main body of GUIDEx, is for those people who design, execute, analyze and report on such experiments. These experimenters are the backbone of the community and should benefit from the full detail of the 14 Principles and 8 Case Studies.

GUIDEx is intended to be a guide for clients, people who ask the questions that lead to experiments and campaigns and for whom reports are prepared. It is also for those who decide how the question will be addressed and approve the methods that will be applied. It is hoped that this pocketbook will act as an introduction to the full GUIDEx and so help stimulate better communication among military officers, government officials and the defense scientific community of the allied nations on all matters associated with defense experimentation.

Paul Labbé
Chair, TTCP JSA AG-12

Table of Contents

Introduction

Increasingly, nations such as the United States, Great Britain, Canada, Australia, New Zealand and indeed NATO itself are relying on experimentation to assist in the development of their future military forces. For example, the United States Department of Defense stresses the importance of experimentation as the process that will determine how best to optimize the effectiveness of its joint force to achieve its vision of the future (US Joint Staff 2000). Is this confidence, in the ability of experimentation to support the military transformation process, appropriate? Certainly, experimentation has proven itself in the sciences and technology by producing dramatic advances. Can the methods of experimentation that have so expeditiously and radically developed science and technology be applied to the military transformation process to achieve similar advances in military effectiveness?

The thesis of this guide is that robust experimentation methods from the sciences can be adapted and applied to military experimentation and will provide the basis for

advancements in military effectiveness in the transformation process. The authors have structured the relevant experimentation material under 14 Principles, which ensure that defense experimentation programs positively impact coalition organizations' ability to evolve force capabilities of the future. Also, they have provided an experimentation-planning flowchart that in one page shows what needs to be done, together with a set of Case Studies that demonstrate the value of the principles in practice.

GUIDEx is not meant to duplicate information already available in other documents and textbooks on experimentation such as those referenced here, [ABCA 2004; Alberts and Hayes 2002, 2005; Dagnelie 2003; Radder 2003; Shadish *et al.* 2002] or on command and control (C2) assessment [NATO 2002], but organizes and expands this detailed information under 14 Principles to guide successful defense experimentation.

Scope

GUIDEx is about the use of the experimental method in the defense domain. A number of terms are used by the TTCP nations to describe such activities, including "warfighting experimentation", "defense experimentation" and "military experimentation". GUIDEx has settled on a single term, *Defense Experimentation* in order to present its ideas in a consistent manner. Defense Experimentation is defined here as "the application of the experimental method[1] to the solution of complex defense capability development problems, potentially across the full spectrum of conflict types[2], such as warfighting, peace-enforcement, humanitarian relief and peace-keeping". GUIDEx also presents the idea of ***Integrated Analysis and Experimentation Campaigns***, in which experiments are combined with other analytical techniques; both to tackle larger problems that would not be possible with single experiments, and to exploit the strengths of different techniques.

[1] The major focus of GUIDEx is experiments based upon field events and human-in-the-loop virtual simulations, but the principles of GUIDEx are also applicable to experiments based on analytic wargames and constructive simulations.

[2] Most of the examples available to this guide have been based on warfighting scenarios, simply because of the legacy of the primary focus of defense experimentation to date.

Experiments and Science

In about 400 B.C., philosophers Socrates and Plato investigated the meaning of knowledge and methods to obtain it using a *rational-deductive* process, or pure logic (logic), without reference to the real world. Aristotle was a transitional figure who advocated observation and classification, bridging to later scientists like Ptolemy and Copernicus who developed *empirical-inductive* methods that focused on precise observations and explanation of the stars. These early scientists were not experimenters. It is only when later scientists began to investigate earthly objects rather than the heavens, that they uncovered a new paradigm for increasing knowledge.

In the early 1600s, Francis Bacon introduced the term experiment and Galileo moved from astronomical observations to conducting earthly experiments by rolling balls down an inclined plane to describe bodies in motion. The realization that manipulating objects would yield knowledge spawned a new research paradigm, one unimagined in the previous 2000 years of exploring the out-of-reach heavens. The basis of this new science paradigm called experimentation (the *empirical-deductive* approach) was a simple question [Feynman 1999]:

"If I do this, what will happen?" The key to understanding experimentation, and the characteristic that separates experimentation from all other research methods, is manipulating something to see what happens. The scientific aspect of experimentation is the manipulation of objects under controlled conditions while taking precise measurements. In its simplest form [Shadish *et al.* 2002: p. 507], an experiment can be defined as a process "*to explore the effects of manipulating a variable.*"

Designing Valid Experiments

Principle 1. Defense experiments are uniquely suited to investigate the cause-and-effect relationships underlying capability development.

Principle 2. Designing effective experiments requires an understanding of the logic of experimentation.

Principle 3. Defense experiments should be designed to meet the four validity requirements.

Improved capabilities cause improved future warfighting effectiveness. Experimentation is the unique scientific method used to establish the cause-and-effect relationship of hypothesized capabilities. If experimenters design the five experiment components to meet the four experiment validity requirements, defined later, the defense experiment will provide the scientific evidence to proceed. Defense experiments are essential to develop empirical- and concept-based capabilities that yield implementable prototypes. The use of a "**develop–experiment–refine**" approach ensures that a rigorous methodology relates new capabilities to warfighting effectiveness. The development and delivery of defense concepts and capabilities is thus supported through experimentation.

Experiment Hypotheses

To understand cause-and-effect relationships between capabilities and increased warfighting effectiveness is to understand experiment hypotheses. Any national or coalition capability problem may be stated as: **Does A cause B?** An experimental capability or concept—a new way of doing business—is examined in experimentation to determine if the proposed capability **A** causes the anticipated military effect **B**. The experiment hypothesis states the causal relationship between the proposed solution and the problem.

Hypothesis

If… **"proposed change"**
Then… **"improved warfighting capability"**

It is an "*If...then...*" statement, with the proposed cause–innovative concept—identified by the *if* clause, and the possible outcome—the problem resolution—identified by the *then* clause.

Components of an Experiment

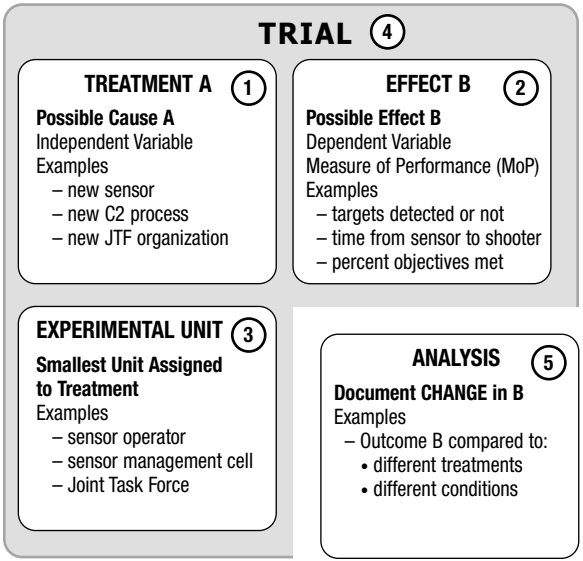
All experiments—large or small, field or laboratory, military or academic, applied or pure—consist of five components[3] [Shadish *et al.* 2002: p. 2]:

1. The *treatment*, the possible cause **A**, is a capability or condition that may influence warfighting effectiveness.
2. The *effect* **B** of the treatment is the result of the trial, an increase or decrease in some measure of warfighting effectiveness.
3. The *experimental unit*[4] executes the possible cause and produces an effect.
4. The *trial* is one observation of the *experimental unit* under treatment **A** or under the alternative **~A** to see if effect **B** occurred, and includes all of the contextual conditions of the experiment.
5. The *analysis* phase of the experiment compares the results of one trial to those of another.

[3] For application of these concepts to test and evaluation, see [Kass 1997].

[4] An experimental unit includes all operators with their gear, procedures, and concept of operations. In experimentation, the apparatus includes the experimental unit and necessary conditions for effecting changes and observing effects.

Five Components of any Experiment



These five components are useful in understanding all defense experiments including large field experiments. Some field experiments are grand exercises with multiple experimental initiatives (possible causes), sometimes as many as 20 to 30 different initiatives in one experiment. To be useful, each individual experimental initiative should be configurable as a unique mini-experiment with its own

subset of the five components. Each initiative is a particular treatment with its own experimental unit (operators in one area of the task force), its own set of outcome measures, and its own set of trial conditions. However, in practice it is very difficult to maintain independence among these many experiments within the large exercise, which makes it difficult to isolate specific causal influences.

What Is a Good Experiment?

A good, or valid, experiment provides information to ascertain whether **A** caused **B** [Shadish *et al.* 2002: p. 3]. **Four logically sequenced requirements** are necessary to achieve a valid experiment.[5] A simple experiment example will illustrate these four requirements. A proposed concept postulates that new sensor capabilities are required to detect future targets. An experiment to examine this proposition might employ *current sensors* on the first day of a two-day experiment and a *new sensor capability* on the second day. The primary measure of effectiveness is the number of targets detected. The experiment hypothesis could be: "If new sensors are employed, then target detections will increase."

[5] Many detailed good practices developed by experiment agencies through experience (and described in recent books such as [Alberts and Hayes 2002, 2005]) can be organized under these four requirements and the 14 Principles.

1 *Ability to use the new capability A*

Developing and generating the new experimental capability for the experiment is often a major resource commitment. In an ideal experiment, operators employ the experimental capability, in this case the new sensors, to its optimal potential; thereby allowing the new capability to succeed or not succeed on its own merits. Unfortunately, this ideal is rarely achieved. A lesson repeatedly learned from defense experiments is that new experimental capabilities are frequently not fully realized in the experiment.

A number of things can go wrong with an experimental surrogate. For example, the hardware or software does not work as advertised or anticipated. The experiment players may be undertrained and not fully familiar with its functionality. Because the experimental treatment represents a new capability, the trial scenario and potential outcomes may not be sensitive to the new capability's enhanced performance.

A valid experiment design ensures that the new capability works under relevant conditions prior to execution, that the operators are adequately trained to employ it appropriately, and that the scenario is sufficiently sensitive to determine the capability's effectiveness. Experimenters continually monitor these aspects during experiment execution. If the experimental sensors **A** do not function during the experiment, the new capability will most likely not affect the military unit's ability to detect targets **B**, which is the next experiment validity requirement.

2 *Ability to detect a change in the effect B*

When the player unit correctly employs a new capability, does it result in any noticeable difference in the effect **B** during the experiment trial? Ideally, a change in the number of detections accompanies a transition from old to new sensors. If this is not the case, this may be because there is too much experimental noise[6]—the ability to detect change is a signal-to-noise ratio problem. Too much experimental error produces too much variability, hampering detection of a change. Reduction of experiment variation, through data collection calibration, limited stimuli presentations, and a controlled external environment, mitigates experiment-induced error. In addition, since the computation of variability in statistics decreases as the number of repetitions increases, a larger sample size increases the signal-to-noise ratio making it easier to detect change.

Analysts measure change in effectiveness by comparing the results of one experiment trial to those of another. Typically, different experiment trials represent different levels of applications of the same capability, alternative competing capabilities, or the same capability under different conditions. A change in military effectiveness may also be detected by comparing the results of an experiment trial to a pre-existing baseline, a task standard, or a desired process.

3 *Ability to isolate the reason for change in the effect B*

If an experimenter employed a useable capability that produced a noticeable increase in the number of target detections, was the observed change in detections due to the intended cause—changing from old sensors to new—or due to something else? In the sensor-experiment example, an alternative explanation for the increase in

[6] Experimental noise interferes with the observation of the desired variable at a required degree of precision.

detections on the second day could be that of a learning effect. That is, the sensor operators may have been more adept at finding targets because of their experience with target presentations on Day One and, consequently, would have increased target detections on Day Two, whether or not different sensors were employed. An increase in operator experience coincidental with a change in sensors would dramatically alter the interpretation of the detected change in effectiveness. An experiment outcome with alternative explanations is a *confounded* result. Scientists have developed experimentation techniques to eliminate alternative explanations of the cause of change: counterbalancing the presentation of stimuli to the experimental unit, the use of placebos, the use of a control group, random assignment of participants to treatment groups, and elimination or control of external influences.

4 *Ability to relate the results to actual operations*

If the player unit ably employed the capability, and if an experimenter detected change and correctly isolated its cause, are the experiment results applicable to the operational forces in actual military operations? The ability to apply, or *generalize*, results beyond the experiment context pertains to experiment realism and robustness. Experiment design issues that support operational realism revolve around the representation of surrogate systems, the use of operational forces as the experimental unit, and the use of operational scenarios with a realistic reactive threat. To ensure the operational robustness, the experiment should examine multiple levels of threat capabilities under various operational conditions.

Experiments during Capability Development and Prototyping

Nations employ a variety of processes to support development of improved empirical- and concept-based capabilities and are, increasingly, employing defense experimentation to support the delivery of this improved warfighting effectiveness. These capability development and prototyping processes are not the same across the different nations (in some nations these processes are referred to as **concept development and experimentation, CD&E**). However, in most cases they follow similar **develop–experiment–refine** stages. For the purposes of GUIDEx, therefore, a generic description of these stages is presented with the hope that the ideals embodied can be mapped onto each nation's own way of doing business.

Stage	Aim
Discovery	To clarify future warfighting problems and to seek potential solutions.
Refinement	To examine and refine the extent to which proposed capabilities or concepts solve military problems.
Assessment	To ensure that solutions from refinement are robust; that they are applicable to a wide range of potential operational requirements in an uncertain future.
Prototype Refinement	To transition capability surrogates into potential operational capabilities by developing complete prototype packages for front line commands.
Prototype Validation	To provide the final demonstrated evidence that the prototype capability can operate within theater and will improve operational effectiveness.

Experiments are required throughout a capability development and prototyping process. They provide an empirical method to explore new capabilities, to refine concepts, and to validate new prototypes for implementation. For example, during *refinement*, experiments quantify the extent to which proposed capabilities solve military problems. Experiments also examine capability redundancies and tradeoffs and reveal capability gaps. Prior *discovery* stage activities only speculate whether proposed further capabilities would solve identified gaps in military effectiveness, whereas experimentation during *refinement* empirically substantiates and quantifies the extent proposed capabilities increase effectiveness in specific case examples. In some instances, experimentation may suggest prototypes for early implementation, or identify areas needing future investigation. Experiments during *assessment*, on the other hand, investigate the robustness of the solution developed during *refinement* for possible future military operations. These experiments examine different future contingencies, different multinational environments, and different threat scenarios to ensure that the *refinement* stage solution is robust; that it is applicable to a wide range of potential operational requirements in an uncertain future.

Prototypes derived from the earlier stages are often not ready for immediate operational use. Experiments during *prototype refinement* can transition concept prototypes

into potential operational capabilities by developing complete prototype packages for front line commands. These experiments develop the detailed tactics, techniques, procedures (TTPs), and organizational structures for the prototype as well as developing the tasks, conditions, and standards to facilitate training. They can also examine the latest hardware and software solutions and their interoperability with existing fielded systems. Experiments during *prototype validation* provide the final demonstrated evidence to the combatant commander that the prototype capability can operate within theater and will improve operations. Often these experiments are embedded within exercises or other training events and are used to validate the predicted gains in effectiveness of the force.

Integrated Analysis and Experimentation Campaigns

Principle 4. **Defense experiments should be integrated into a coherent campaign of activities to maximize their utility.**

Principle 5. **An iterative process of problem formulation, analysis and experimentation is critical to accumulate knowledge and validity within a campaign.**

Principle 6. **Campaigns should be designed to integrate all three scientific methods of knowledge generation (studies, observations and experiments).**

Principle 7. **Multiple methods are necessary within a campaign in order to accumulate validity across the four requirements.**

Experimentation is a necessary tool in addressing large capability development problems, but this should be embedded in an integrated campaign of experiments, studies and analytical activities. Such Integrated Analysis and Experimentation Campaigns would typically also have an integrated analytical and management process, and use a variety of techniques to ensure that weaknesses in one technique can be mitigated by others.

Campaigns use a mix of defense experiments and parallel studies to understand the problem's context, the associated warfighting concept and the capabilities required. The product of the campaign is advice to decisionmakers on the utility, versatility and maturity of the concept and the capabilities required to implement the concept. Campaigns can address issues at all levels from joint and combined operations to platforms and components.

An integrated campaign using a variety of techniques ensures that weaknesses in one technique can be mitigated by others. Where results (*e.g.*, inferences) correlate between activities, it increases confidence and where they diverge, it provides guidance for further investigation. It is only when all activities are brought together in a coherent manner and the insights synthesized, that the overall problem under investigation is advanced as a whole.

Such campaigns can address force development issues at any level, for example: technological (*e.g.*, systems of systems), tactical, operational, as well as strategic. Instances of activities at each of these levels in Australia, for example, are as follows:

- at the **technological** level: helicopter operations within a combined arms team, surface and sub-surface platforms for maritime operations, and the JSF within the air control system;
- at the **tactical** level: amphibious and airmobile task groups;
- at the **operational** level: the capability balance required to achieve the Future Warfighting Concept; and finally,
- at the **strategic** level: the Effects Based Operations concept is being developed in conjunction with many government agencies.

Why use a Campaign

An integrated analysis and experimentation campaign will be required for a variety of reasons. There may be resource or political reasons why a campaign is preferred to a single activity, or more likely it will be necessary because without a coordinated campaign, the problem or issue under investigation simply cannot be satisfactorily resolved. A campaign allows the problem to be approached in a coordinated, manageable manner with a variety of analytical techniques and allows a degree of iteration and synthesis between activities that help ensure

that the overall problem is sufficiently addressed. The problem may initially be ill-defined and a campaign of activities will allow assessment and adjustment as the problem is refined. Some of the analytical reasons for using a campaign approach are described in the following sub-sections.

- **Problem Characteristics.** Military capability development problems are generally complex and coercive. The socio-technical nature of the system and the interaction between the components and the environment characterize the system as complex. The importance of an opposing force, itself a socio-technical system, means the system is coercive. Many problems that might be explored through defense experimentation are simply too complex to be dealt with in a single activity.
- **Increased Confidence**. An integrated campaign of experiments and other activities allows a gradual build-up of the knowledge surrounding the problem or issue under investigation, leading to a more refined and robust concept. This increases confidence that the findings are valid and creates a systematic body of knowledge to inform and investigate capability development.
- **Synthesis of Military and Analytical Skills.** A campaign, by integrating different techniques, provides improved opportunity for analytical and military skills to be applied to the problem.
- **Problem Formulation.** When the strategic environment is uncertain and unprecedented, and the impact of technology unknown, the experience base is usually too narrow to conduct the problem formulation confidently. Within the campaign we must therefore build a **synthetic experience base** and the process of scientific inquiry is used to increase our confidence in the problem formulation.

Iterating Methods and Experiments

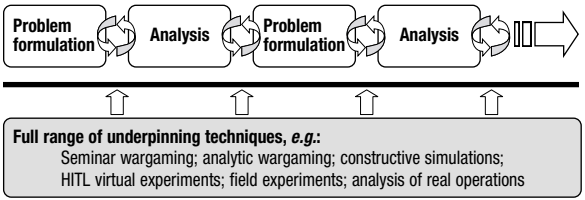
The initial stage of any campaign is problem formulation. Effective problem formulation is fundamental to the success of all analyses, but particularly at the campaign level because the problems are normally ill-defined, complex and adversarial, involving many dimensions and a rich context. Problem formulation involves decomposition of the military and analytical aspects of the problem into appropriate dimensions. Decomposition cannot normally be achieved without detailed analysis using a matrix of tools such as seminars and defense

experiments supported by analytical studies and operational experience. Detailed analysis also assists in the reconstruction of the problem segments and interpretation of results.

In dealing with fuzzy or uncertain interactions, the problem formulation process needs to explore and understand the significance of each interaction before making (or seeking from customers) assumptions about it. This involves keeping an open mind, during the early stages of problem formulation, about where the boundaries lie and their dimensional nature. This is difficult because it makes the process of modeling the problem more complicated. A call for hard specification too early in that process must be avoided. In the end, of course, the problem must be formulated in order to solve it, but formulation should be an output from the first full iteration, not an early input to it.

As shown in the following illustration, the problem is being formulated and refined throughout the entire campaign in an iterative cycle that never really completes until the campaign itself completes. The process of problem formulation and analysis is undergoing constant review to reshape the direction of the campaign and to ensure that the real issue or concept is being addressed.

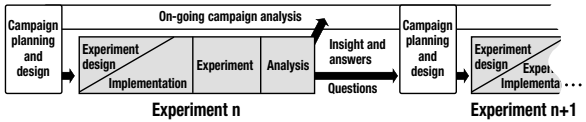
Coherent Management and Communication Framework



Wargames, and in particular seminar wargames, have an important role in problem formulation. In wargaming it is possible to balance the physical and psychological aspects of the problem by using warfighters as the players while adjudicating their actions using models or rulesets. Most importantly, wargaming introduces an adversary early in the problem formulation process, providing a stressful environment in which to explore the concept and develop the hypotheses for subsequent analysis. Although human-in-the-loop simulations and live simulations also introduce a human adversary, they are frequently too expensive and unwieldy for the problem formulation phase.

Integration of Scientific Methods

The aim of a campaign is to integrate a range of methods: experiments (observations with manipulation—empirical-deductive); observational studies (observations without manipulation—empirical-inductive) and analytical studies (rational-deductive) into a coherent package that addresses a complex capability development problem. The phases of campaign design are the same as for any evaluation, that is, problem formulation and analysis. The complexity arises because after the completion of each activity the problem formulation is reassessed and adjusted and subsequent activities may be redesigned. As a result a campaign plan is a flexible instrument, with a supporting risk-management framework and an iterative approach to constantly review and reshape the remainder of the campaign to ensure that the overall goals are achieved.



In all likelihood, seminars, workshops, historical analysis, and the like, will also be required as part of the campaign to support and help inform the experimenters who will ultimately address the overall question. The campaign plan process must take these other activities into account within its design phase. The ultimate aim is to synthesize the outputs from all activities into coherent advice to the decisionmakers.

Different Methods Offer Different Strengths

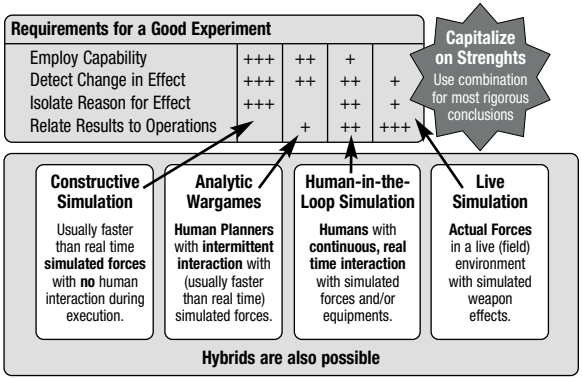
All experiments must strike a balance among the four experiment validity requirements. Attempts to satisfy one work against satisfying the other three. Consequently, 100 percent-valid experiments are unachievable. Precision and control increase the ability to detect change and to isolate its cause, but decrease the ability to apply the results to imprecise, real-world situations. Experiments designed to identify change emphasize strict control of trial conditions and feature multiple repetitions of similar events; experiments designed to relate results emphasize free-play, uncertainty, and reactive threats. Each individual experiment design must consider requirement tradeoffs in order to minimize the loss of one requirement due to the priority of another.

Most defense experiments use some form of simulation, which can be grouped into one of four general methods: *constructive simulation, analytic wargames, human-in the-loop simulation, and live (field) simulation*. Each of these four methods has its own strengths and weaknesses with respect to the four experiment validity requirements discussed previously. Since one particular method cannot satisfy all four requirements, an integrated analysis and experiment campaign requires multiple methods.

Constructive simulations are those in which no human intervention occurs in the play after designers choose the initial parameters and then start and finish the simulation. Constructive simulations are a mainstay of military analytical agencies. They allow repeated replay of the same battle under identical conditions, while systematically varying parameters—the insertion of a new weapon or sensor characteristic, the employment of a different resource or tactic, or the encounter of a different threat. Experiments using constructive simulations with multiple runs are ideal to detect change and to isolate its cause. Because modeling complex events requires many assumptions, including those of variable human behavior, critics often question the applicability of constructive simulation results to operational situations.

Analytic wargames typically employ command and staff officers to plan and execute a military operation. At certain decision points, the Blue players give their course of action to a neutral, White cell, which then allows the Red players to plan a counter move, and so on. The White cell adjudicates each move, using a simulation to help determine the outcome. A typical analytic wargame might involve fighting the same campaign twice, using different capabilities each time. The strength of such wargames for experimentation resides in the ability to detect any change in the outcome, given major differences in the strategies used. Additionally, to the extent that operational scenarios are used and actual military units are players, analytic wargames may reflect real-world possibilities. A major limitation is the inability to isolate the true cause of change because of the myriad differences found in attempting to play two different campaigns against a similar reactive threat.

Rigorous experimentation requires multiple methods to meet the four validity requirements.



Human-in-the-loop simulations represent a broad category of real-time simulations with which humans can interact. In a human-in-the-loop defense experiment, military subjects receive real-time inputs from the simulation; make real-time decisions, and direct simulated forces or platforms against simulated threat forces. The use of actual military operators and staffs allows this type of experiment to reflect warfighting decisionmaking better than experiments using purely constructive simulation. However, when humans make decisions, variability increases, and changes are more difficult to detect and consequently to attribute to the cause.

Live simulation is conducted in the actual environment, with actual military units and equipment and with operational prototypes. Usually only weapon effects are actually simulated. As such, the results of experiments in these environments, often referred to as field experiments, are highly applicable to real situations. Good field experiments, like good military exercises, are the closest thing to real military operations. A dominant consideration however, is the difficulty in isolating the true cause of any detected change since field experiments include much of the uncertainty, variability, and challenges of actual operations; but they are seldom replicated due to costs.

Different Methods during Capability Development and Prototyping

As potential capabilities advance through capability development and prototyping stages, the following considerations are useful in selecting which of the four experiment validity requirements to emphasize. For example, finding an initial set of potential capabilities that empirically show promise is most important in the *refinement* stage. Experiments in this early stage examine idealized capabilities (future capabilities with projected characteristics) to determine if they lead to increased

effectiveness, and are dependent on the simulation-supported experiment, using techniques such as constructive simulation, analytic wargames and human-in-the-loop simulation. Accurately isolating the reason for change is not critical at that stage, as the purpose is only to apply a coarse filter to the set of idealized capabilities. However, during the *assessment* stage, quantifying operational improvements and correctly identifying the responsible capabilities is paramount in providing evidence for concept acceptance. This is also dependent on experiments with better-defined capabilities across multiple realistic environments. Experiments conducted using constructive simulation can provide statistical defensible evidence of improvements across a wide range of conditions. Human-in-the-loop and field experiments with realistic prototypes in realistic operational environment can provide early evidence for capability usability and relevance. Early incorporation of the human decisionmaker in this way is essential, as the human operators tend to find new ways to solve problems.

In *prototype refinement* experiments, one should anticipate large effects, otherwise its implementation might not be cost effective. Accordingly, the experiment can focus on the usability of working prototypes in a realistic experiment environment. Isolating the real cause of change is still critical when improving prototypes. The experiment must be able to isolate the contributions of training, user characteristics, scenario, software, and operational procedures. As previously described, human-in-the-loop and field experiments provide the opportunity for human decisionmakers to influence development. In *prototype validation*, human decisionmakers ensure that the new technology can be employed effectively. Prototype validation experiments are often embedded within joint exercises and operations.

Employing Multiple Methods to Increase Rigor

Since experiments using the four main simulation methods emphasize the four validity requirements differently, an integrated analysis and experimentation campaign must capitalize on the strengths of each method to accumulate validity. For example, the model–exercise–model paradigm integrates the strengths of, on the one hand, the constructive simulation (*i.e.*, "model") and, on the other, any of the methods that involve human interaction (*i.e.*, "exercise" in a generic sense). This technique is especially useful when resource constraints

prohibit conducting side-by-side baseline and alternative comparisons during wargames and field experiments.

In the model-exercise-model paradigm, the early experiments using constructive simulation examine multiple, alternative, Blue-force capability configurations and baselines. Analysis of this pre-exercise simulation allows experimenters to determine the most beneficial Blue-force configuration for different Red-force scenarios. An analytic wargame, human-in-the-loop or field experiment can then be designed and conducted, which provides independent and reactive Blue- and Red-force decisionmakers and operators. One can then re-examine this optimal configuration and scenario.

Experimenters use the results of the exercise to calibrate the original constructive simulation for further post-event simulation analysis. Calibration involves the adjustment of the simulation inputs and parameters to match the simulation results to those of the experiment, thus adding credibility to the simulation. Correspondingly, rerunning the pre-exercise alternatives in the calibrated model provides a more credible interpretation of any new differences observed in the simulation. Additionally, the post-exercise calibrated simulation improves analysts' ability to understand fully the implications of the experiment results by conducting "what if" sensitivity simulation runs. Experimenters examine what might have occurred if the Red or Blue forces had made different decisions during the experiment.

The model–exercise–model method increases overall experiment validity by combining the contrasting strengths of the following methods:

1. experiments using constructive simulation, which is strong in detecting differences among alternative treatments, and
2. experiments using either human-in-the-loop simulation, analytic wargame, or field experiments, which are stronger in incorporating human decisions that better reflect the actual operating environment.

This paradigm also helps to optimize operational resources by focusing the exercise event on the most critical scenario for useful results, and by maximizing the understanding of the event results through post-event sensitivity analysis.

Considerations for Successful Experimentation

Principle 8. Human variability in defense experimentation requires additional experiment design considerations.

Principle 9. Defense experiments conducted during collective training and operational test and evaluation require additional experiment design considerations.

Principle 10. Appropriate exploitation of modeling and simulation is critical to successful experimentation.

Principle 11. An effective experimentation control regime is essential to successful experimentation.

Principle 12. A successful experiment depends upon a comprehensive data analysis and collection plan.

Principle 13. Defense experiment design must consider relevant ethical, environmental, political, multinational, and security issues.

Principle 14. Frequent communication with stakeholders is critical to successful experimentation.

This guide identifies a number of considerations that are intended to support the practical implementation of experiments. These considerations relate to the need to recognize and accommodate the human element in experiment design, and they also provide advice on how to make the best use of operational test and evaluation events or training exercises. They also give guidance on some issues relating to modeling and simulation, on the implementation of good experiment control and highlight national regulations, security rules and practices that may need special consideration; and finally, there are also some practical steps that can be taken to achieve good communications.

Human Variability

The implications arising from using human subjects in defense experimentation are often overlooked. Most, if not all defense experiments examine impacts on socio-technical systems but experiment designs rarely cater sufficiently for the human element. Because humans are unique, highly variable and adaptable in their response to an experimental challenge, they are more than likely to introduce a large experimental variability. In addition, humans will have different experiential baselines in terms of, for example training and aptitude and, unlike technology, will become tired and possibly demotivated.

They may also learn during experiments. The experiment design and the data analysis and collection plan must recognize and accommodate human variability, which will be much larger than would be predicted if the socio-technical system were treated solely as technology. What is sometimes overlooked is that this variability provides important information on why a socio-technical system responds to a challenge in a particular way. Indeed there is an argument that human variability should not be minimized, as this would lose important information. High variability may indicate a fault in the system under examination, or in the experiment design. An understanding of the impact of human variability on experiment design and outcome is a fundamental skill required by all experimenters.

Regardless of the experimenter's ability to control human variability, it is important, if possible, to measure it. This is done mainly to see if detected effects can be explained in terms of human variability rather than the experimental treatments. For example, where a single group is the subject for all the treatments, then learning by that group during and between the treatments may have a confounding effect on the whole experiment. It may be possible to measure learning effects within each treatment, and thus estimate any confounding effect of learning between treatments. Of course, this may increase the complexity of the experiment design as the data analysis will then also need to control for human variability measures and assess their impact upon the main variables.

Although objective measures of variables are favored by experimenters, subjective measures are important for ascertaining the mental processes underlying observed behaviors. This information may be important, especially if a subject adapts to using a capability in a way not considered by the experimenter. Asking subjects why they have changed their behavior can enhance understanding of maladaptive ways of using of a new capability. Consideration needs to be given to the timing of subjective interviews, particularly whether they should take place soon after the action occurs or at the end of the experiment. The former may be obtrusive to the subjects and may impact the results, with the latter being affected by factors such as memory decay and motivation.

Exploiting Operational Test and Evaluation and Collective Training Events

Opportunities to conduct experimentation may be found in training exercises and in operational test and evaluation (OT&E) events. Operational assessments, in particular, provide an opportunity for conducting experimentation with substantial technological and expert staff support. The drive to conduct experimentation activities during training exercises and OT&E events is almost entirely due to the difficulty of acquiring the resources (equipment, estate, human) to undertake defense experiments of any significant size. Arguably, the equipment programs that require most support from experimentation are those intended to enhance *collective* rather than team or individual effectiveness, and thus collective groups of personnel (which may comprise command teams with higher and lower controllers) are required to undertake that experimentation. It is a simple fact of life in the early 21st Century that most nations generally do not have units and formations available to dedicate to experimentation, except for the most limited-scale activities. Therefore exploiting routine training exercises and other collective events should be given serious consideration.

Exploiting collective training (exercises) has a range of benefits as well as disadvantages and a variety of factors must be taken into account in both planning and execution. The principal one is that training always has primacy and the experimenter has little control over events, thus the skill is in understanding the constraints that the exercise opportunity will present and knowing how to work within them. Exploiting training exercises for the purposes of experimentation is most achievable during the prototype validation phase of a capability development program when functional prototypes exist.

The potential to include experimentation within OT&E programs is very high. This is so in part because many of the components of OT&E events are the same as their counterparts in experiments. They are well supported by the technical/engineering community and valued by the operational community as a component of the operational readiness process. The operational community will therefore generally be engaged in OT&E events and the potential to include experiments in these events as well can be very good. An important benefit to experimenters is the OT&E infrastructure, which includes engineering/technical staffs and facilities; planning support; test support during execution and evaluation

support for the after-action review or report (AAR). The benefit from the use of OT&E staffs and facilities is realized because of the strong overlap between the two processes. An important benefit to the OT&E community is that the prototypes from experiments may soon be operational systems. In such circumstances, there is a significant advantage to be obtained by the inclusion of OT&E staffs in experimentation on these systems.

Although training exercises and OT&E events do not allow execution of elaborate experiment designs because it would impede training and impact operational readiness, scientific methodology and the four experiment validity requirements can be applied to such embedded experiments. Experimentation in these situations naturally provides the strongest venue for meeting the fourth experiment validity requirement, *i.e.*, the ability to relate results to actual operations. While operational necessity restricts the ability to meet the first three experiment validity requirements in training exercises, and to a lesser extent in OT&E events, the experimenter can ameliorate the limitations to some degree. With respect to the first experiment validity requirement, *i.e.*, the ability to use the new capability, prototype testing prior to the training exercise enhances the usability of the experimental capability and should ensure that it will function correctly during the exercise trials. This is less of an issue for OT&E, as this activity is generally for validating the performance of new operational systems and the testing is implicit. Additionally, to address the second experiment validity requirement in training exercises, *i.e.*, the ability to detect a change in the effect, establishing a pre-exercise definition of expected performance and comparing the prototype's actual performance during the exercise to its expected performance provides the necessary ability to detect change. For OT&E, the performance of new operational systems is typically documented in manuals and validated computer models may exist. Therefore, the baseline system performance should be well established and the potential for detecting change should be good.

While the ability to isolate the reason for the observed change effect, *i.e.*, the third experiment validity requirement, is the most problematic in experimentation embedded in training exercises, experimenters can nevertheless achieve some level of satisfaction here as well. When examining different capabilities during a single exercise, the experimenter should conduct different

prototype trials at different times so the effects of one prototype do not influence the effects of the other. It is prudent to have an experienced exercise "observer-controller" view the prototype trial to assess the extent that any observed results were the results of the experimental capability instead of unintended causes. Additionally, showing that the rigorous experiment data accumulated during the concept development phase of the prototype is still relevant to the exercise conditions also supports GUIDEx third experiment validity requirement. Experimentation embedded in OT&E events also creates considerable challenges for meeting the third experiment validity requirement. The best approach in this case is through comprehensive, detailed data collection, which is typically the case in OT&E events anyway.

Finally, for both the use of training exercises and OT&E events, a Model-Exercise-Model paradigm that was successfully calibrated to the event results would allow follow-on sensitivity analysis to demonstrate that inclusion and exclusion of the experimental capability accounted for decisive simulation differences.

Training Exercises

Benefits

- Availability of experimental subjects in large numbers
- High level of engagement of experimental subjects
- Use of training infrastructure
- Moderate sample sizes, for repeated exercise series
- Ability to use repeated exercises as a control group, or baseline
- They rate highly in terms of relating any detected change to real operations.

Constraints

- Exercises are designed to stimulate various training points that may not satisfy an experiment design
- Training has primacy—can a genuine experiment design be fitted around training?
- Scenarios and settings designed for training purposes
- Limited opportunities to make intrusive changes to the exercise or collected data intrusively
- Can results be published without breaching the anonymity of the training audience?
- Interventions by Exercise Control for training reasons, *e.g.*, the training force is winning too easily
- Exploitation of an exercise too early in a unit's training cycle can yield poor results, *e.g.*, the collective skills may be too low.

OT&E Events

Benefits

- Availability of operational staff and platforms
- High level of engagement of technical community
- Use of OT&E infrastructure
- Moderate sample sizes, for repeated test series
- Ability to use repeated tests as a control group, or baseline
- Strong potential for relating any detected change to real operations.

Constraints

- OT&E events are designed to quantify aspects of equipment performance or to determine if a standard is being met that may not satisfy an experiment design
- OT&E has priority and the experiment may not interfere with test objectives
- Scenarios and settings designed for OT&E purposes
- Limited opportunities to make intrusive changes to the test or collected data intrusively
- Can results be published without breaching the anonymity of the test audience?

Modeling and Simulation Considerations

This guide presents modeling and simulation (M&S) as intrinsic to conducting most defense experiments. There is now a wide range of M&S techniques available and this makes the innovative use of M&S cost effective for many defense experimentation applications. However, there are some significant issues associated with selecting both the types of M&S to be used and the specific elements of the experiment federation.

A balanced view of fidelity and validity

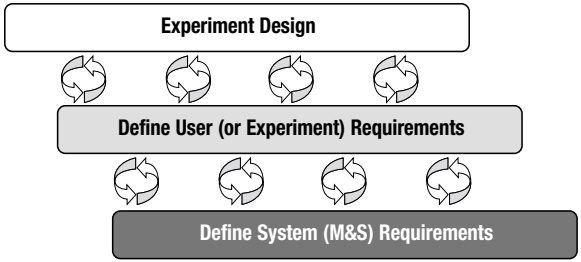
For many years, as rapidly increasing computing power led to many new modeling possibilities, there was a generally held view that greater fidelity, or accuracy, was always better. Indeed, many took the term "validity" to be almost synonymous with fidelity and detail. The modern view is that validity actually means "fit for purpose," with the *purpose* being to execute the desired experiment design. This means that we should consider the main measure of merit for M&S to be *adequacy*, not *fidelity*. The experiment design should effectively define what level of fidelity is adequate. Furthermore, the main point of modeling is to rationalize the complexity of real life by simplifying it. In "The Lanchester[7] Legacy" [Bowen and McNaught 1996: Vol. III, Ch. 9], the authors wrote: "It has long been understood by Operational Analysts that, in dealing with complicated situations, simple models that provide useful insights are often to be preferred to models that get so close to the real world that the mysteries they intend to unravel are repeated in the model and remain mysteries." We can therefore imply an axiom that M&S should be as simple as possible while remaining adequate for the task in hand.

M&S definition

It is a key principle that the definition of the M&S to be used in an experiment should be derived from the experiment design, and not the other way around. However, rarely will practitioners have the luxury of completing their experiment design and then moving through a user requirements and subsequently system requirements definition process in sequence. Usually a concurrent process is necessary, with the processes beginning in the order given above. A spiral development process can then take place. There are several well-

[7] F.W.Lanchester was one of the pioneers of military operational research.

established processes for achieving this, *e.g.*, the US Federation Development and Execution Process (FEDEP) and the European Synthetic Environment Development and Exploitation Process (SEDEP).



Modeling the process to be examined by the experiment

Experiments and observational studies (where a concept is subjected to objective observation, but without manipulation) are intrinsically connected to the idea of hypotheses. The hypothesis is simply a plausible proposition about either causal or associative relationships. Thus in a general sense there is always implicitly a model of the process being experimented with by virtue of there being one or more hypotheses. However, it is possible, and in most cases desirable, to model the process in advance in a much more tangible way, regardless of whether a strict model-exercise-model paradigm is being followed. In particular, architectural frameworks such as Zachman [Zachman 1987] and DoDAF[8] represent an excellent and increasingly popular means to describe military problems and potential candidate solutions in a variety of different ways. When a **model-exercise-model** paradigm is being followed, process models based on these frameworks can often be preferable to complex constructive combat simulations.

Experiment Control

Experimentation is intrinsically a controlled activity, although the degree of possible and required control varies from case to case. The experiment design should be explicit in describing which variables must be controlled in order to prevent rival explanations for the findings, and which variables can be allowed to remain uncontrolled though usually recorded. It should also describe the control regimes to be put in place to ensure that this occurs in practice. The identification of intervening variables and learning effects must be well understood. However, simply outlining the required measures in the

[8] DoD Architecture Framework, see [DoDAF Working Group 2004]

experiment design document is not sufficient. The experiment director and his team must actively seek to impose the required controls throughout the planning and execution phases of the experiment.

Experiment Design

The experiment design process is a logical journey from the questions to be answered, or hypotheses to be tested, to the detailed definition of the experiment. Thus the experiment design is the cornerstone of the control regime throughout the life of the experiment, since it sets out in broad terms what needs to be done. Success in designing experiments is rooted in early stakeholder engagement to establish objectives and intent. An integrated analysis and experimentation campaign goes a long way toward providing the framework for detailed stakeholder guidance. Furthermore, nothing allows for the control of variables during experiment design more than early, firm decisionmaking. The longer decisions on scenario, participation, funding, technical environment, and study issues are allowed to linger, the more options the experiment designers must keep open and the harder it is to control the variables that can affect the outcome of the experiment.

Experiment Planning

The planning of major defense experiments requires a management team, which takes the decisions required to settle high-level issues, has oversight on the activities of the various teams, and ensures that the experiment planning and organization develops toward the objectives in a timely manner. A series of reviews throughout the planning period is usually necessary to ensure that the process of preparing for the experiment is remaining on track. For larger experiments, *e.g.*, joint or coalition ones, it is common to employ conferences for this purpose, organized and run by the management team; typically three or four might be used.

Experiment Execution

The experiment management team usually transforms into the control staff during execution. The controller's role is to ensure that the experiment is progressing according to schedule or to be on top of the situation if it is not. The controller observes the players and collects their input daily and works closely with the analysts in monitoring the progress of the experiment. The controller provides feedback to the experiment director and implements changes as required to ensure the event achieves the

experiment objectives. In doing so, the controller must deal with military judgment (observations from the players) and scientific objectivity (input from the analysts).

Experiment Analysis

The analysis or assessment team for an experiment should ideally be derived at least partly from the experiment design team, and they should work closely with the team responsible for the concept under experiment and the team responsible for providing the experiment's technical environment. Initially, they should review the concept and approach planned to conduct the experiment and prepare an analysis plan to meet the needs of the experiment design. During the course of an experiment, analysts compare observations and results and begin to integrate their views of what is being learned from the experiment. As sufficient data is collected, analysts begin to form preliminary insights. However, the temptation to announce some startling finding (especially one that it is believed the experiment sponsor will like) should be resisted at all costs, because it is quite likely that when the analysis is complete, that finding will at best need to be modified, and at worst, changed altogether. Thus, first impressions should generally be conservative; this is an important control consideration.

Data Analysis and Collection

Data collection is designed to support the experiment analysis objectives that in turn rely on a conceptual model underlying the experiment. The data analysis offers the opportunity to revisit the underlying conceptual model identified for the experiment and determines cause-and-effect relationships. A data analysis and collection plan is an essential part of an experiment.

A significant part of the experiment consists of gathering data and information. Interpreting the information into findings and combining them with already known information to obtain new insights tends to be challenging. Once it is determined what needs to be measured, a decision is required to identify the data necessary and to analyze it using appropriate (usually statistical) analysis techniques. The plan ensures appropriate and valid data are generated and that the key issues of the experiment are addressed. When determining analytical techniques to use, an estimate for the number of observations must be considered, depending on the expected variability in the dependent variables and the number of them. It is essential to prioritize and ensure there are sufficient observations for all objectives, measures of performance, and measures of effectiveness requiring analysis. There

exist various types of collection mechanisms used in experiments.

Questionnaires (also referred to as surveys) are often used in data collection. They can be used to gather numerous types of information. The participants' background can be obtained through this means. This can be done before the start of the experiment. The participants can also be questioned about aspects of the experiment such as their perceptions about the systems and processes tested, their view on others participating, strengths and weaknesses of the systems and processes as well as recommended improvements.

With information systems becoming more crucial, *Automated Collection Systems* to collect data are now more important. It is important to determine what clock each system that is used to collect data is synchronized to in order to facilitate analysis.

Observers have an important part in the experiment by capturing interactions between participants. For instance they take notes about what is going on, crucial events taking place, notable behaviors and other such activities. Observers can also be used to provide a chronological narrative of the events that occurred. This provides documentation about what happened during the experiment and can be used to explain why certain results occurred.

Ethics, Security and National Issues

This guide describes a number of different aspects of defense experimentation. However, in addition, distinctive national regulations, security rules and practices should not be underestimated and proper consideration must be given to them in planning experiments.

Environmental considerations

Wherever there is live activity, there will be some level of environmental impact. In particular, great care must be taken regarding proximity to historical or cultural sites. As well as legal and multinational environment issues, environmental constraints generally will have an impact on the scope of any live experiment or exercise. It is essential that results be interpreted in the light of all environmentally imposed artificialities. The test and training communities have been working with environmental issues for years and there is no reason for the experimentation community to deviate from the various protocols that already exist.

Security considerations

Even within single-nation experiments, security issues can give rise to real practical problems. In particular, the rise of secure digital command, control, communications, computers and intelligence (C4I) and sensitive intelligence, surveillance, target acquisition and reconnaissance (ISTAR) sources (which are often themselves at the centre of the experiment purpose) has resulted in security considerations becoming much more prominent in the design and execution of defense experiments than hitherto. As a general rule, the lower the security classification of these elements, the lower the cost and risk of the experiment and thus experiments should be run at the lowest classification level possible. This is not to say, of course, that undue efforts should be made to make everything unclassified or artificially low in classification. As previously discussed, all experiments are compromises, and the experimenter needs to decide where the benefits of (for example) higher classification and therefore higher fidelity representations of equipments or scenarios outweigh the benefits of using lower classification analogues.

Ethics considerations

Any experiment, which involves human subjects and human data collectors, could potentially pose ethical issues. By recruiting subjects to undertake an experiment, or by exposing the data collector to a potentially hazardous military environment the experimenter is expecting them to operate outside their normal working practices. Although ethics is a complex field, its fundamental concerns in professional contexts can be defined. Research that lacks integrity is considered to be ethically unacceptable, as it not only misrepresents what it claims to be but also misuses resources. In addition, there is an obligation for defense experiments to comply with relevant national Health and Safety legislation and to provide working conditions that would ensure, as far as reasonably practicable, a healthy and safe working environment for experimenters and subjects alike.

Communication with Stakeholders

The final product of any defense experiment must be the evidence that the right question has been addressed and the evidence required for its findings to be exploited effectively. This will also provide the experimenter with the necessary foundation for advising on the applicability and feasibility of advancing an evaluated concept, or elements of a concept, toward eventual realization as actual operational capabilities. Good and continuous communication is central to achieving such a successful

outcome; and yet it is still possible to find an experiment, or integrated analysis and experimentation campaign, which does not have a rational plan for communicating with stakeholders.[9] A communications plan must consider how the different stages in running an experiment may require different approaches to good communication; stages such as determining the right set of questions and issues to be addressed, maintaining the confidence of key stakeholders that the potential changes to their priorities are being considered, ensuring all stakeholders have appropriate access during the experiment and making sure that they understand the output.

Determining the right set of question and issues

A key prerequisite to a single experiment or campaign is the identification of the origins of the question to be addressed and identification and commitment of key stakeholders. One difficulty is that the obvious stakeholder is often not the person that originally posed the question. Therefore an initial step must be to chase down the origins of the question, and from that define the key stakeholders who need to be influenced. However, the question may arise from many sources and it may not always be possible to directly engage or even identify the original source. For example the question may have arisen from a strategic plan, which states that "there is a need to enhance interoperability with our allies to a level which will allow us to undertake concurrent medium scale operations." This will reflect a political imperative, and whoever is responsible for the strategic plan may have appointed intermediaries whose task is to implement this directive. In this case, these are all key stakeholders, and it is essential to determine their relationships and how they work together. Intermediaries will have formed their own understanding of the question being posed and defined a campaign to implement the directive.

Communicating in the run up to the experiment

Although this will be a particularly busy period, it is essential that regular dialogue be maintained with the stakeholder community prior to the experiment. By maintaining this regular dialogue, changes in priorities can be quickly identified and accommodated.

[9] Stakeholders are defined as persons who have a vested interest in the product from the experiment or campaign.

Communicating during the experiment

In most cases the major interaction with stakeholders occurs during the visitor day. Visitors should be encouraged to view the entire experimentation process from the pre-brief to the post exercise wash up, and invited to observe and interact with the subjects in a way that does not interfere with the experiment. Additional attendance outside the specific visitor day of stakeholders with a direct involvement in the campaign implementation improves communication in that they are then briefed at regular intervals.

Communicating after the experiment

A well-written report will contain a one-page abstract, an executive summary and a full report. The traditional approach to dissemination of results has been to produce a paper that is sent to key stakeholders, with or without a presentation. While this has obvious merits the general experience is that this approach tends to produce "shelf-ware."[10] It should be remembered that these are busy people who will wish to gain quick appreciation of the key issues and findings, in order to exploit the information.

[10] A UK term, which means that the report is produced but never read in full.

A far better approach is to continue the dialogue with the key stakeholders to determine how the work has been received, to assist in interpreting results and, more importantly, to advise on how it should be exploited. Where the experiment is part of a wider campaign supporting concept or capability development, the experimenter may also have the opportunity to advise on the consequences for the over-arching concept of the particular experiment findings.

GUIDEx Experiment and Campaign Planning Flowchart

In order to help practitioners in applying the GUIDEx principles to address their specific problems, the following flowchart was developed. This is by no means a prescriptive recipe for perfect experimentation, but an attempt to lay out the chronological sequence for experiment and campaign related activities and to show the iterations and linkages between various stages of the experimentation process. Indeed, GUIDEx encourages that the specific application of Principles to a given problem should be tailored according to the scale and nature of the issue under investigation. There is no single "best" way to undertake experimentation, rather the skill of the practitioner is to use a degree of artistic license in applying the science advocated within GUIDEx in order to maximize what can be achieved for a given problem under real-world constraints of resources, time, expectation and understanding.

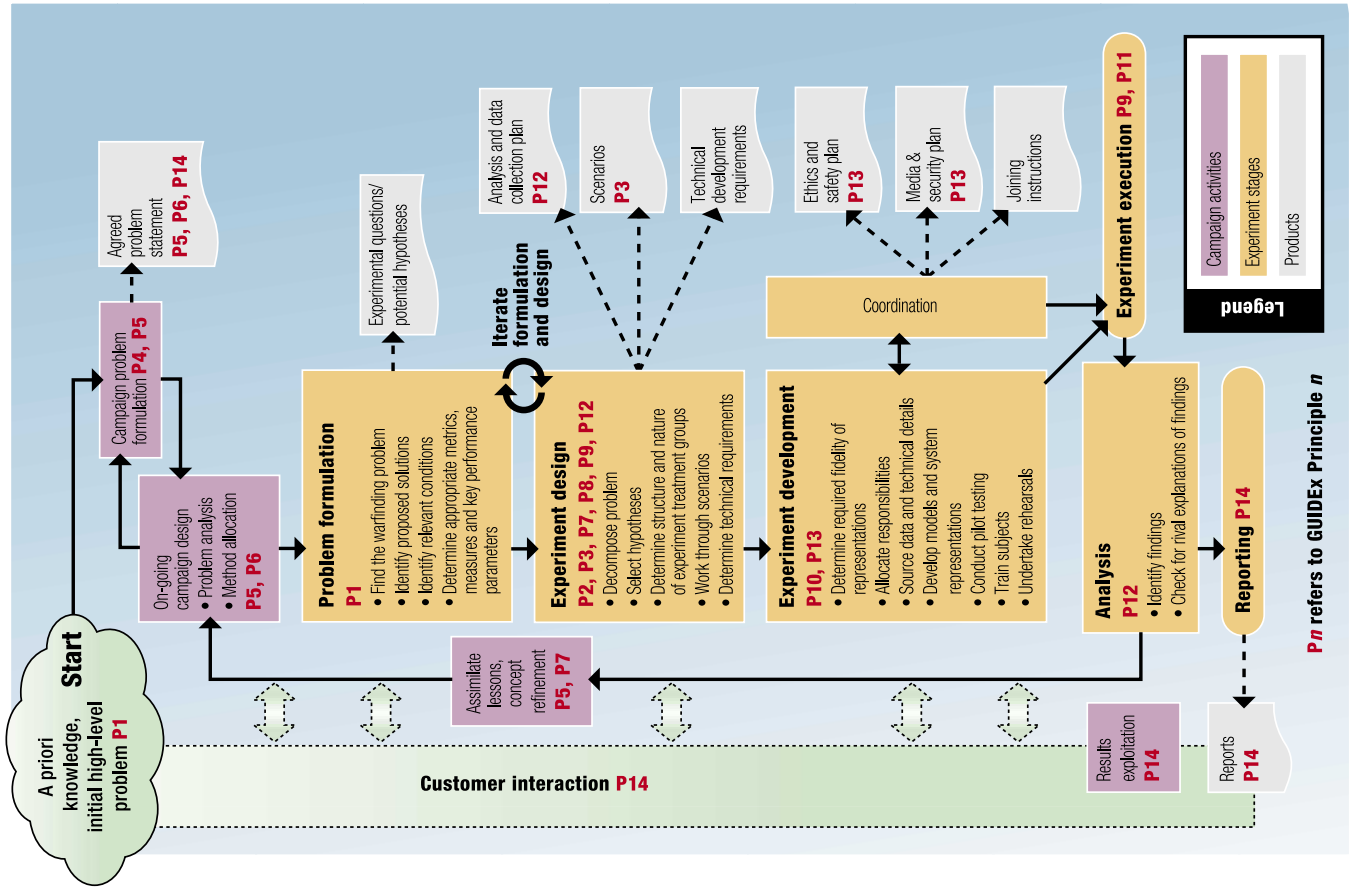
The color code of the flowchart separates the integrated analysis and experimentation campaign activities (in purple) from the specific individual experiment stages (in orange). The grey areas indicate the products of the experimentation process, while green shows the customer or stakeholder interactions. The flowchart itself begins from the green cloud at the top-left hand corner, representing the initial problem, as posed by the customer.

The campaign of integrated analysis and experimentation then commences with a number of iterations around the campaign problem formulation and campaign design loop in order to develop with the customer an agreed campaign-level problem statement. During this process the campaign designer begins to identify the analytical methods and experiments that might be used to answer the problem. Once a required experiment is identified, the more detailed process of experiment problem formulation can begin. Again, the flowchart suggests that the problem formulation should iterate and overlap with the experiment design in order to ascertain the problem scope suitability for experimentation imposed by real-world considerations. A number of potential experimental questions may require some initial design work to be undertaken before an acceptable, workable and useful problem defined can then be submitted to a complete experiment design and development. The lesson is "be prepared for exploratory

activities or false starts before one can move forward with a good concept for detailed design."

The flowchart outlines some of the products needed for successful experimentation, such as analysis and data collection plans, technical development requirements, ethics and safety plans and finally joining instructions for the participants. The practitioner's role at this stage is to manage the competing demands of technical development, customer and player expectation, legislative requirements, rehearsal and training requirements while still maintaining overall control of the scientific and analytical rigor. Finally the experiment itself is executed and the process of analysis and reporting can begin.

In general as the individual experiment is being planned, designed and undertaken, the campaign analysis continues and once the results from the experiment emerge from the collected data, the campaign itself may evolve to take account of the knowledge gained. Lessons must be assimilated. If necessary, further experimentation or analytical activities can be undertaken and the cycle repeats. Throughout this entire process, the interaction with the customer is key to ensuring that the answers generated do indeed answer the questions posed.



21 Threats to a Good Defense Experiment

5 Experiment Components	① Ability to **Use** Capability	② Ability to **Detect** Change	③ Ability to **Isolate** Reason for Change: Single Group	③ Ability to **Isolate** Reason for Change: Multiple Groups	④ Ability to **Relate** Results to Operations
① Treatment	**1.** Capability not workable: Do the hardware and software work?	**5.** Capability variability: Are systems (hardware and software) in use in like trials the same?	**11.** Capability changes over time: Are there system (hardware or software) or process changes during the test?	N/A	**18.** Nonrepresentative capability: Is the experimental surrogate functionally representative?
② Players	**2.** Player non-use: Do players have the training and tactics, techniques and procedures (TTPs) to use the capability?	**6.** Player variability: Do individual operators/units in like trials have similar characteristics?	**12.** Player changes over time: Will the player unit change over time?	**15.** Player differences: Are there differences between groups unrelated to the treatment?	**19.** Nonrepresentative players: Is the player unit similar to the intended operational unit?
③ Effect	**3.** No potential effect in output: Is the output sensitive to capability use?	**7.** Data collection variability: Is there a large error variability in the data collection process?	**13.** Data collection changes over time: Are there changes in instrumentation or manual data collection during the experiment?	**16.** Data collection differences: Are there potential data collection differences between treatment groups?	**20.** Nonrepresentative measures: Do the performance measures reflect the desired operational outcome?
④ Trial	**4.** Capability not exercised: Do the scenario and Master Scenario Event List (MSEL) call for capability use?	**8.** Trial conditions variability: Are there uncontrolled or unmonitored changes in trial conditions for like trials? Look for intervening variables not recorded.	**14.** Trial conditions change over time: Are there changes in the trial conditions (such as weather, light, start conditions, and threat) during the experiment?	**17.** Trial conditions differences: Are the trial conditions similar for each treatment group?	**21.** Nonrepresentative scenario: Are the Blue, Green, and Red conditions realistic?
⑤ Analysis	N/A	**9.** Low statistical power: Is the analysis efficient and the sample sufficient? **10.** Violation of statistical assumptions: Are the correct analysis techniques used and error rate reduced?	• The purpose of an experiment is to verify that **A** causes **B**. • A valid experiment allows the conclusion, **A** causes **B**, to be based on evidence and sound reasoning... — by reducing or eliminating the 21 known threats to validity.		

4 Experiment Requirements

21 Threats to a Good Experiment

Building on the work of Cook and Campbell [Cook and Campbell 1979], one can identify the things that can go wrong in an experiment. Cook and Campbell call these **threats to validity**, in other words, they are identified problem areas that can cause one to not meet any one of the four experiment requirements presented from page 10 to 15 of this Pocketbook. While Cook and Campbell identified 33 threats to validity, they have been combined and distilled down to 21 potential threats to defense experiments. Moreover, they have been rearranged into a two-dimensional matrix to better systematically illustrate how the threats to experiment validity can be understood and treated with respect to each of the four requirements and the five experiment components. Additionally, many names of their threats to validity have been changed to reflect military experiment terminology. For example, learning effects is the substitute of Cook and Campbell's *maturation*.

All good experiment practices are then ways to eliminate, control, or ameliorate these threats. A good experiment plan would show how each has been accounted for and countered.

The two-dimensional framework of this Table provides a substantial advantage over the traditional "laundry list" of good practices. The framework associates different good practices with each of the four experiment requirements. This facilitates understanding why particular good practices are important and the impact on experiment validity if the threat is not properly attended to. For example, it is impossible to implement all of the good practices in any particular experiment. Thus, an understanding of the impact of unimplemented good practices is critical to designing the "best available" experiment. Furthermore, associating good practices with the different experiment components allows the experiment designer to see the interaction of good practices across all aspects of the experiment. Fortunately, when developing an experimentation campaign, one can achieve a higher level of fulfillment of the good practices by using the particular power of complementary experimentation approaches.

GUIDEx Case Studies

The following is a high-level overview of the results of the eight Case Studies offered by GUIDEx.

1. **Testing Causal Hypotheses on Effective Warfighting:** This was a series of experiments for a common operational picture (COP) experimental treatment condition using a Persian Gulf air/sea scenario where all parties—higher echelon and lower echelon—had both the national intelligence supported big picture and the local tactical picture. This combination was experimentally proven to be superior technology for such operations, resulting in greater shared situation awareness and better bottom line combat effectiveness.

2. **UK Battlegroup Level UAV Effectiveness:** This experiment supported a major UK unmanned air vehicle (UAV) acquisition program in demonstrating the huge information gathering potential of UAVs at the tactical level, compared to existing ISTAR assets. However, equally importantly, it showed that if integration into the supported HQs is not achieved effectively, then the resulting information overload can have a hugely detrimental effect on mission success.

3. **UK NITEworks ISTAR Experiment:** The UK, like other nations, is presently investing heavily in ISTAR sensors and systems. However, it is widely recognized that effective information requirements management (IRM) is vital to the efficient use of those systems. This experiment investigated both technological and procedural means of improving IRM. It showed conclusively that a collaborative working environment with appropriate working practices would have a major beneficial effect on IRM effectiveness. This assisted the development of ISTAR management priorities in the UK.

4. **Pacific Littoral ISR UAV Experiment (PLIX):** This Case Study provides insights difficult to capture without experimentation; the strong hypothesis of identifying and tracking all targets proved not to be attainable even though sensor coverage was nominally complete, pointing to integration requirements for an effective ISR architecture.

5. **An Integrated Analysis and Experimentation Campaign:** Army 21 / Restructuring the Army 1995-99: This campaign demonstrated the importance of detailed problem definition and an iterative approach based on wargaming, field trials and analytical studies. The warfighting concept under test was found to fail under realistic environmental constraints. However, the results led to an alternative concept, which is the basis for current Australian Army force development.

6. **The Peregrine Series:** a Campaign Approach to Doctrine and TTP Development: This on-going campaign of experiments and studies is contributing directly to the development of the doctrine for employment of the Australian Army's new Armed Reconnaissance Helicopters and demonstrates how experimentation can be used to inform capability development questions at unit level and below.

7. **Multinational Experiment Three (MNE 3):** Despite the complexity of the MNE 3 effects-based planning (EBP) experiment and the findings that the concept and supporting tools require further development, the event demonstrated the potential for EBP to make a coalition task force a more effective instrument of power. It also showed the benefits for collaboration to produce the best ideas from a collective thought process in a coalition, which included a civilian interagency component.

8. **Improved Instruments Increase Campaign Values:** While improved experimentation instruments provided the opportunity to generalize some results, they also increased the validity of campaign's results and knowledge generation synthesized for future information management systems.

Epilogue

The thesis of GUIDEx is that, while it is true that defense experiments are not like some highly abstracted and inanimate laboratory experiments, the logic of science and experimentation can be applied to defense experiments to produce credible tests of causal claims for developing effective defense capabilities. An overview of that thesis has been presented in this pocketbook version of GUIDEx.

This guide presents experimentation practices and examples resulting from the deliberation of the AG-12 participants, who have all had experience in their own countries' defense experimentation efforts. The reader is encouraged to apply and adapt the 14 Principles laid out in GUIDEx to improve experimentation across the TTCP nations, although they do not express national positions. Many examples within the guide are based on the specific perspective and experience of different lead-nation authors with contributions from other participants: they may require supplementary effort to relate them to national perspectives. It is anticipated that as GUIDEx is used, practitioners will develop additional good practices and examples, and this will stimulate an update to GUIDEx in the future.

Acronyms, Initialisms and Abbreviations

AAR	after-action review or report
ABCA	American, British, Canadian, Australian Armies
ACT	Allied Command Transformation (NATO)
AG	Action Group
AU	Australia
C2	command and control
C4I	command, control, communications, computers and intelligence
CA	Canada
CCRP	Command and Control Research Program
CD&E or CDE	concept development and experimentation
CFEC	Canadian Forces Experimentation Centre
COBP	code of best practice
COP	common operational picture
CPX	command post exercise
CS	Case Study (With capitals for GUIDEx CSs, case study otherwise)
DISA	Defense Information Systems Agency
DoD	Department of Defense
DRDC	Defence Research and Development Canada
Dstl	Defence science and technology laboratory
DSTO	Defence Science and Technology Organisation
EBP	effects-based planning
GCCS	Global Command and Control System
GIG	Global Information Grid
GUIDEx	TTCP Guide for Understanding and Interpreting Defense Experimentation
HITL	human-in-the-loop
HQ	headquarters
HUM	TTCP Human Resources and Performance Group
HW/SW	hardware/software
IRM	information requirements management
ISR	intelligence, surveillance and reconnaissance
ISTAR	intelligence, surveillance, target acquisition and reconnaissance
JFCOM	Joint Forces Command
JSF	Joint Strike Fighter
JTF	Joint Task Force
MAR	TTCP Maritime Systems Group
MBM	model-based-measures
M-E-M	model-exercise-model
MNE	Multinational Experiment
MoD	Ministry of Defence (UK)
MoE	measure of effectiveness
MoM	measure of merit
MoP	measure of performance
MSEL	master scenario event list

NAMRAD	Non-Atomic Military Research and Development
NATO	North Atlantic Treaty Organisation
NCO	network centric operations
NCW	network centric warfare
NEC	network enabled capability
NITEworks	Network Integration Test and Experimentation works
NL	National Leader
OT&E	operational test and evaluation
OTH-T	over-the-horizon targeting
PLIX	Pacific Littoral ISR Experiment
TP	Technical Panel
TRADOC	US Army Training and Doctrine
TTCP	The Technical Cooperation Program
TTPs	tactics, techniques and procedures
TUAV	tactical unmanned air vehicle
UAV	unmanned air vehicle
UK	United Kingdom
US/USA	United States of America
USV	uninhabited surface vehicle

References

ABCA. 2004. “American, British, Canadian, and Australian Armies' Standardization Program Analysis Handbook (draft for unlimited distribution).” 66 p. http://abca.hqda.pentagon.mil/

Alberts, D.S. and R.E. Hayes. 2002. *Code of Best Practice for Experimentation. Washington*, DC: CCRP. 436 p.

—. 2005. *Campaigns of Experimentation; Pathways to Innovation and Transformation*. Washington, DC: CCRP. 227 p.

Bowen, C. and K.R. McNaught. 1996. “Mathematics in Warfare: Lanchester Theory.” p. 141-156 in *The Lanchester Legacy, Volume 3 - A Celebration of Genius*, edited by N. Fletcher. Coventry, UK: Coventry University Press.

Cook, Thomas D. and Donald T. Campbell. 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Boston: Houghton Mifflin. 405 p.

Dagnelie, Pierre. 2003. Principes d'expérimentation: planification des expériences et analyse de leurs résultats: Electronic edition. 397 p. http://www.dagnelie.be

DoDAF Working Group. 2004. “DoD Architecture Framework, Version 1.0.” p. 87.

Feynman, R.P. 1999. *The Meaning of It All: Thoughts of a Citizen Scientist*. USA: Perseus Books Group. 133 p.

Kass, R.A. 1997. “Design of Valid Operational Tests.” *International Journal of Test and Evaluation* (June/July):51-59.

Radder, H. 2003. The Philosophy of Scientific Experimentation. Pittsburgh, PA: University of Pittsburgh Press. 311 p.

Shadish, W.R., T.D. Cook, and D.T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin. 623 p.

Thomke, S.H. 2003. *Experimentation Matters; Unlocking the Potential of New Technologies for Innovation*. Boston: Harvard Business School Press. 307 p.

US Joint Staff. 2000. “Joint Vision 2020.” edited by US Department of Defense: US Government Printing Office.

Zachman, J.A. 1987. “A Framework for Information Architecture.” *IBM System Journal* 26(3):276-292.

Index

Acknowledgements

The preparation of this document would not have been possible without selected collaborative activities conducted under the TTCP umbrella that included meetings, conferences and workshops with participation from JSA, HUM and MAR (group, technical panel and action group members), interactions with ABCA, NATO RTO and ACT (Allied Command Transformation), and the direct and indirect contributions by participating country experts.

The participants of TTCP JSA AG-12 with the collaboration of several experts produced this document. They are listed below by alphabetical order of family name.

Bowley, Dean	Defence Science & Technology Organization (DSTO)	AU
Comeau, Paul	Canadian Forces Experimentation Center (CFEC)	CA
Edwards, Dr Roland, NL[11]	Defence Science and Technology Laboratory (Dstl)	UK
Hiniker, Dr Paul J.	Defense Information Systems Agency (DISA)	US
Howes, Dr Geoff, NL	Defence Science and Technology Laboratory (Dstl)	UK
Kass, Dr Richard A., NL	Joint Forces Command (JFCOM), Experimentation	US
Labbé, Paul, Chair	Defence Research & Development Canada	CA
Morris, Chris	NITEworks	UK
Nunes-Vaz, Dr Rick	Defence Science & Technology Organization (DSTO)	AU
Vaughan, Dr Jon, NL	Defence Science & Technology Organization (DSTO)	AU
Villeneuve, Sophie	Canadian Forces Experimentation Center (CFEC)	CA
Wahl, Mike	Joint Forces Command (JFCOM), Experimentation	US
Wheaton, Dr Kendall, NL	Canadian Forces Experimentation Center (CFEC)	CA
Wilmer, Col Mike	US Army Training and Doctrine (TRADOC)	US

[11] UK National Leader until June 2004, then Dr Geoff Howes joined AG-12 as UK NL.

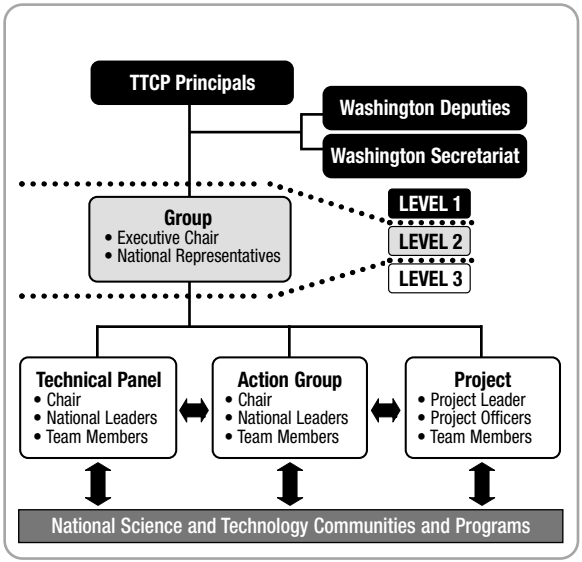
An electronic copy of this report is available at the following URL: **http://www.dtic.mil/ttcp**

National reviewers: AU, Dr Paul Gaertner; CA, Chris McMillan, UK, George Pickburn; and US, Dr Paul Hiniker.

Copy editor: France Crochetière

TTCP Document Feedback

The aim of TTCP is to foster cooperation within the science and technology areas needed for conventional (*i.e.*, non-atomic) national defense. The purpose is to enhance national defense and reduce costs. To do this, it provides a formal framework that scientists and technologists can use to share information among one another in a quick and easy fashion. Its structure is illustrated below:



More information on TTCP can be found on its public Website at **http://www.dtic.mil/ttcp/**

For the purpose of maintaining and updating TTCP unlimited distribution documents (publications that, due to their value to the academic, scientific and technological communities, are widely distributed) readers and users of these documents are invited to email their appreciation, comments and suggestions for future editions to **ttcp_docfeedback@dtic.mil**

This address is administered by the TTCP Washington Staff, who will pass feedback onto the appropriate document point of contact. For more information on TTCP document feedback, please see the TTCP guidance document 'POPNAMRAD', which can be found on the public website.

Notes